ChemMedChem 2018 vol.13 N6, pages 540-554

Mapping of the Available Chemical Space versus the Chemical Universe of Lead-Like Compounds

Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

© 2018 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim This is, to our knowledge, the most comprehensive analysis to date based on generative topographic mapping (GTM) of fragment-like chemical space (40 million molecules with no more than 17 heavy atoms, both from the theoretically enumerated GDB-17 and real-world PubChem/ChEMBL databases). The challenge was to prove that a robust map of fragment-like chemical space can actually be built, in spite of a limited (≪105) maximal number of compounds (“frame set”) usable for fitting the GTM manifold. An evolutionary map building strategy has been updated with a “coverage check” step, which discards manifolds failing to accommodate compounds outside the frame set. The evolved map has a good propensity to separate actives from inactives for more than 20 external structure–activity sets. It was proven to properly accommodate the entire collection of 40 m compounds. Next, it served as a library comparison tool to highlight biases of real-world molecules (PubChem and ChEMBL) versus the universe of all possible species represented by FDB-17, a fragment-like subset of GDB-17 containing 10 million molecules. Specific patterns, proper to some libraries and absent from others (diversity holes), were highlighted.

http://dx.doi.org/10.1002/cmdc.201700561

Keywords

computer chemistry, generative topographic mapping, library comparison, molecular diversity, structure analysis

References

[1] D. Horvath, Methods Mol. Biol. 2011, 672, 261–298.
[2] S. J. Lusher, R. McGuire, R. C. van Schaik, C. D. Nicholson, J. de Vlieg, Drug Discovery Today 2014, 19, 859–868.
[3] I. V. Tetko, BIGCHEM—Big Data in Chemistry, 2016, http://bigchem.eu.
[4] M. Johnson, G. M. Maggiora, Concepts and Applications of Molecular Similarity, Wiley, New York, 1990.
[5] A. Schuffenhauer, P. Ertl, S. Wetzel, M. A. Koch, H. Waldmann, J. Chem. Inf. Model. 2007, 47, 47–58.
[6] T. Varin, A. Schuffenhauer, P. Ertl, S. Renner, J. Chem. Inf. Model. 2011, 51, 1528–1538.
[7] B. Zhang, M. Vogt, G. M. Maggiora, J. Bajorath, J. Comput.-Aided Drug Des. 2015, 29, 937–950.
[8] I. T. Jolliffe, Principal Component Analysis, Springer, New York, 2002;
[9] G. H. Dunteman, Principal Components Analysis, Sage, Newbury Park, CA, 1989.
[10] D. Horvath, M. Lisurek, B. Rupp, R. Kühne, E. Specker, J. von Kries, D. Rognan, C. D. Andersson, F. Almqvist, M. Elofsson, P.-A. Enqvist, A.-L. Gustavsson, N. Remez, J. Mestres, G. Marcou, A. Varnek, M. Hibert, J. Quintana, R. Frank, ChemMedChem 2014, 9, 2309–2326.
[11] T. Kohonen, Self-Organizing Maps, Springer, Heidelberg, 2001.

[12] C. M. Bishop, M. Svensén, C. K. I. Williams, Neurocomputing 1998, 21, 203–224.

[13] C. M. Bishop, M. Svensén, C. K. Williams, Neural Comput. 1998, 10, 215–234.

[14] H. A. Gaspar, I. I. Baskin, G. Marcou, D. Horvath, A. Varnek, J. Chem. Inf. Model. 2014, 55, 84–94.

[15] H. Gaspar, G. Marcou, D. Horvath, A. Arault, S. Lozano, P. Vayer, A. Varnek, J. Chem. Inf. Model. 2013, 53, 3318–3325.

[16] N. Fechner, G. Papadatos, D. Evans, J. R. Morphy, S. C. Brewerton, D. Thorner, M. Bodkin, Bioinformatics 2013, 29, 523–524.

[17] P. Sidorov, H. Gaspar, G. Marcou, A. Varnek, D. Horvath, J. Comput.-Aided Mol. Des. 2015, 29, 1087–1108.

[18] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, Nucleic Acids Res. 2012, 40, D1100–D1107.

[19] L. Ruddigkeit, R. van Deursen, L. C. Blum, J.-L. Reymond, J. Chem. Inf. Model. 2012, 52, 2864–2875.

[20] PubChem Ver. 2010, US National Institutes of Health, https://pubchem.ncbi.nlm.nih.gov/.

[21] ISIDA, Laboratoire de Chemoinformatique, Strasbourg, France, 2012, http://infochim.u-strasbg.fr/spip.php?rubrique41.

[22] F. Ruggiu, G. Marcou, A. Varnek, D. Horvath, Mol. Inf. 2010, 29, 855–868.

[23] A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, C. Gaudin, P. Vayer, V. Solov´ev, F. Hoonakker, I. V. Tetko, G. Marcou, Curr. Comput.-Aided Drug Des. 2008, 4, 191–198.

[24] ChemAxon Ver. 2008, ChemAxon, Budapest (Hungary), 2008, https://chemaxon.com/products/chemical-structure-representation-toolkit.

[25] ChemAxon Ver. 2007, ChemAxon, Budapest (Hungary), 2007, https://docs.chemaxon.com/display/docs/Tautomer+Generation+Plugin.

[26] ChemAxon Ver. 2013, ChemAxon, Budapest (Hungary), 2013, https://docs.chemaxon.com/display/docs/pKa+Plugin.

[27] K. Heikamp, J. Bajorath, J. Chem. Inf. Model. 2013, 53, 791–801.

[28] R. Visini, M. Awale, J.-L. Reymond, J. Chem. Inf. Model. 2017, 57, 700–709.

[29] W. A. Warr, Mol. Inf. 2014, 33, 469–476.

[30] D. Horvath, J. Brown, G. Marcou, A. Varnek, Challenges 2014, 5, 450–472.

[31] MontrealQCMOE, Ver. 2015.10, Chemical Computing Group, (Canada), 2015.

[32] A. T. Hagler, E. Huler, S. Lifson, J. Am. Chem. Soc. 1974, 96, 5319–5327.

[33] H. A. Gaspar, P. Sidorov, D. Horvath, I. I. Baskin, G. Marcou, A. Varnek in Frontiers in Molecular Design and Chemical Information Science—Herman Skolnik Award Symposium 2015: Jürgen Bajorath, Vol. 1222 (Eds.: R. J. Bienstock, V. Shanmugasundaram, J. Bajorath), American Chemical Society, Washington, DC, 2016, pp. 211–241;

[34] H. A. Gaspar, I. I. Baskin, G. Marcou, D. Horvath, A. Varnek, Mol. Inf. 2015, 34, 348–356.

[35] K. Klimenko, G. Marcou, D. Horvath, A. Varnek, J. Chem. Inf. Model. 2016, 56, 1438–1454.

[36] P. Tino, I. Nabney, IEEE Transactions on Pattern Analysis and Machine Intelligence 2002, 24, 639–656.

[37] Y. Hu, D. Stumpfe, J. Bajorath, J. Med. Chem. 2016, 59, 4062–4076.

[38] F. Lovering, J. Bikker, C. Humblet, J. Med. Chem. 2009, 52, 6752–6756.